**Descriptive Statistics**

The probability stuff from before is concerned with working out the consequences of knowing the probability distribution of a population, i.e., working out probabilities of measurements given that information. Statistics seems to want to work backwards, to determine what we may conclude about the population, given knowledge of certain measurements.

**Sampling**

A complete description of a population would comprise the information collected from every individual in that population. But that’s usually not feasible, so we collect a sample from the population. We need the sample to be representative.

Nominal data, e.g., gender, can’t be ordered/ranked.

Ordinal data can be ordered/ranked. It can be discrete (categorical) or continuous.

Random sampling is choosing objects at random, yeah.

Systematic sampling is choosing every nth object.

Stratified sampling is breaking the population down into mutually exclusive categories (like male/female/whatever) and randomly sampling from those strata.

Random sampling is important, for instance, in A/B testing, where you have two sample populations which you vary by one metric, A , and test for a commensurate difference in another metric, B. If there is a covariance, you’d like to impute a causal relationship between A and B. But this can only be done if are no confounding variables, i.e., variables that changed with A between the two populations. Confounding variables can theoretically be eliminated by using random samples. That way all differences between the groups are washed out, except for the difference in A.

There are a few ways to organize a study for A/B testing. The best way is a controlled double blind experiment. In this way we randomly assign people to the two groups. And we blind the groups to the knowledge that they differing in A, and we also blind the researches to the knowledge of which group is differing in A. The former can be done by, say, giving placebo pills to one group and real pills to the other. And the latter can be accomplished by keeping even the researchers from knowing whether they’re delivering a real or placebo pill to the two groups. This goes a long way towards eliminating bias.

Often studies cannot be controlled in this fashion – we cannot assign people to smoking vs. non-smoking groups. So we have to do studies where people have already self-selected themselves into these groups. Since the assignment was not random, this enhances the possibility of there being confounding variables present between the groups. Perhaps, in this case, one would attempt a stratified sample of each group to try to mitigate any possible biases.

One more distinction to mention. Say we wanted to examine how age affects sperm count or something. We could follow a person along as they age, and take regular measurements. If we do this for a bunch of random people, starting at the same age, then perhaps we will eliminate all confounding variables. This approach would be called a longitudinal study. Another approach is to gather a random sample of people of all ages, and take measurements, grouped by age. This might work, but could run into confounding variables in that each older age group would be perhaps have had different generational eating habbits, say. For instance it could be that each generation eats better than the previous, and so a decline in sperm counts could just be due to the older generations’ worsening eating habits.

Knowing every piece of data is not especially useful. Usually we are intereted in just a few quantities of note. Some of these are explicated below.

**Histograms**

Histograms are basically discrete probability distributions. And we could construct one from our sample results.

A graph of a number of hours

Description automatically generated

We might want to summarize the data in our histogram. We can do this with measures of center and spread.

**Measures of Center**

If we were to attempt to describe the entire distribution by a single point, we’d have a couple options. These are:

**Mean** – the average.



**Median** - If all data points are listed in numerical order, then the median is the one in the middle, or, if it’s even number of data points (so that there is no single data point in the middle), then it’s the average of the two middles. Median have the nice property that they are less influenced by outliers. So if there are a lot of outliers, the median might be a better representation of the ‘typical’ value than the average is.

**Mode** - The most frequent value. Note there can be multiple values which occur with the same frequency. In this case the sample is called multi-modal.

Say we had a triangular distribution, defined over the interval (0, xmax). A normalized linear probability distribution would be:



What is the mean?



What is the median? It’s the spot with the same number of data points to the left as to the right.



So xmed > xmean. And xmode would be xmax of course. Anyway, can say that xmed will stick closer to the most probabie values than xmean will. xmean is more influenced by outliers than xmed, due to the fact that all samples are not weighted evenly vis a vis xmean, but have the weighting factor ‘x’.

**Measures of Spread**

And then we want ways to quantify the spread of the data. A few of these are:

**Standard Deviation** - Yeah, the population and sample standard deviations are given below. The sample standard deviation is the best estimate for the population standard deviation



where μ is the population mean and is the sample mean. We could go into a little more detail.

**Quartiles** – Quartiles are kind of a generalization of the median. If you have N data points, arranged in numerical order. The quartiles, Q0, Q1, Q2, Q3, Q4 are the values below which 0%, 25%, 50%, 75%, 100% of the data lie. A little more specifically, I guess you could say Q0, Q2, Q4 are the min, median, and max data points. Q1 would be the median of all data points less than Q2, and Q3 would be the median of all data points greater than Q2. The interquartile range (IQR) comprises the data points between Q1 and Q3.

**Percentile** - So the 84th percentile would be the data point behind which 84% of the data lies. I guess you can say that the percentile function is kind of like the cumulative distribution function.

Some graphs illustrate this below.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

So there are different ways to summarize data. One way is called the 5NS (five number summary)



(of course we could say min = Q0 and max = Q4) When graphing a whole bunch of points which themselves have probability distribution functions, a nice way to convey the distribution is with a box plot. Pretty self-explanatory. Note that ‘point just below’, ‘point just above’, refers to an actual data point which is ‘just below’, ‘just above’. The extra red dots above and below are actual data points, that lie outside (Q1 – 1.5\*IQR, Q3 + 1.5\*IQR) interval. And FYI, IQR (‘interquartile range’) = Q3 – Q1. I’m guessing *all* the data points which lie outside the range are displayed? These points are considered outliers. Note the 5NS is conveyed by the box plot. The whiskers are extra.

Chart, diagram, box and whisker chart

Description automatically generated

Another diagram,

Chart, box and whisker chart

Description automatically generated

**Point Estimators**

Guess I’m going to put this here. Say we have some random variable, X with average μ and standard deviation σ. If we don’t know what these are a priori, we can estimate them by means of sampling. Having collected n measurements xi, we might suppose that a good estimate for these is:



But how can we know? Well we look at the random variable behind this point estimator and check to see whether its expectation is equal to the thing we are hoping to estimate. Apropos , we’d consider the random variable,



And now can show that it average is μ,



Since its average is μ, it is said to be non-biased. Other non-biased estimators exist. For instance, could use anything of the sort,



But the advantage of using the one where αi = 1 is that it has the smallest variance. We can calculate the variance,



which is smaller than the variance of any ´ with any other αi’s. Now let’s look at the sample variance random variable, S2,



and calculate its expectation,



So we can see that if we defined the sample variance with 1/n factor, instead of 1/(n-1), our estimate would clearly be off! Similarly, if trying to estimate covariance from a sample, it seems the best predictor is defined kind of like how the standard deviation is, with division by N-1 rather than by N. So we’d say (sum is over all coordinates (xi,yi)):



And sample correlation would be:



Cor is often denoted as r, and is the same thing as R (as in R2 = 1 – SSEf/SSEm) for linear regression. It’s basically an average of correlated z-variables.



So yeah.

**Appendix**

Just curious. Say we have some alternative definition of the variance, σ´2



where μ´ is just some random number, not necessarily the mean. Can we say that σ´2 is smallest when μ´ = μ, the mean? Well let’s consider,



Yes. So interestingly, calculating the variance of a distribution about any point will be larger than calculating it about its mean. So we can formulate a statistic for estimating the population mean by forming the SSE (sum of the squared errors) and finding the μ´ which minimizes it. This is similar to how we find formulas for the linear regression parameters.